Salience-Guided Iterative Asymmetric Mutual Hashing for Fast Person Re-Identification

Cairong Zhao, Yuanpeng Tu, Zhihui Lai, *Member, IEEE*, Fumin Shen, Heng Tao Shen, *Senior Member, IEEE*, and Duoqian Miao

***Abstract*— Person Re-identification (ReID) aims to retrieve the pedestrian with the same identity across different views. Existing studies mainly focus on improving accuracy, while ignoring their efficiency. Recently, several hash based methods have been proposed. Despite their improvement in efficiency, there still exists an unacceptable gap in accuracy between these methods and real-valued ones. Besides, few attempts have been made to simultaneously explicitly reduce redundancy and improve discrimination of hash codes, especially for short ones. Integrating Mutual learning may be a possible solution to reach this goal. However, it fails to utilize the complementary effect of teacher and student models. Additionally, it will degrade the performance of teacher models by treating two models equally. To address these issues, we propose a salience-guided iterative asymmetric mutual hashing (SIAMH) to achieve high-quality hash code generation and fast feature extraction. Specifically, a salience-guided self-distillation branch (SSB) is proposed to enable SIAMH to generate hash codes based on salience regions, thus explicitly reducing the redundancy between codes. Moreover, a novel iterative asymmetric mutual training strategy (IAMT) is proposed to alleviate drawbacks of common mutual learning, which can continuously refine the discriminative regions for SSB and extract regularized dark knowledge for two models as well. Extensive experiment results on five widely used datasets demonstrate the superiority of the proposed method in efficiency and accuracy when compared with existing state-of-the-art hashing and real-valued approaches. The code is released at https://github.com/Vill-Lab/SIAMH.**

***Index Terms*— Person re-identification, hashing, self distillation, mutual learning.**

Manuscript received December 26, 2020; revised June 29, 2021 and August 2, 2021; accepted August 27, 2021. Date of publication September 8, 2021; date of current version September 14, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62076184, Grant 61673299, Grant 61976160, and Grant 61573255; in part by the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education; in part by Shanghai Innovation Action Project of Science and Technology under Grant 20511100700; and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weiyao Lin. *(Cairong Zhao and Yuanpeng Tu contributed equally to this work.) (Corresponding author: Cairong Zhao.)*

Cairong Zhao, Yuanpeng Tu, and Duoqian Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai 200092, China (e-mail: zhaocairong@tongji.edu.cn).

Zhihui Lai is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518060, China (e-mail: lai_zhi_hui@163.com).

Fumin Shen and Heng Tao Shen are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: fumin.shen@gmail.com; shenhengtao@hotmail.com).

Digital Object Identifier 10.1109/TIP.2021.3109508

I. Introduction

PERSON Re-Identification (ReID), refers to the work of retrieving the correct target from a large collection of person images, which has received increasing attention due to its great potential in real-world applications. Tremendous progress has been achieved in the last few years. However, it remains a challenging task owing to various factors, such as viewpoints, severe occlusions and posture variations. To address these issues, existing person ReID methods [1]–[9] mainly focus on performing robust identity-invariable representation learning by facilitating prior knowledge or designing sophisticated frameworks. Most of these methods mainly adopt complex models and extract high-dimensional real-valued features to pursue high performance. However, high computational cost comes with these issues, making the most state-of-the-art models unsuitable for real-world applications, such as security surveillance. Moreover, several large-scale ReID benchmark datasets have been proposed to simulate real-world scenarios recently, which contain a large number of gallery data. It will be very time-consuming for existing real-valued approaches to retrieve the target person in these datasets.

Recently, hashing [11]–[22] has been proved to be a promising way for large-scale image retrieval and applied in a wide range of real-world scenarios. By mapping high-dimensional features into compact binary codes, they can perform fast retrieval in the low-dimensional Hamming space. To guarantee the even distribution and discrimination of hash codes, Lin *et al.* [18] propose an unsupervised hash network. Lai *et al.* [19] introduce a simultaneous feature learning framework that aims to preserve the similarity during the projection. Inspired by them, several supervised hashing ReID methods [23]–[27] have been proposed, which bring remarkable reduction to the memory cost and the query time. However, due to the large information loss in the binarization process, there still exists a large gap between hashing methods and real-valued methods in terms of accuracy. Two main crucial factors can be taken into account for explaining the gap: a) Existing approaches mainly adopt simple frameworks for fast feature extraction, leading to low-quality real-valued features. Therefore, it limits further boosts in accuracy as well, b) To generate discriminative hash codes, they mainly focus on minimizing quantization loss, while ignoring redundancy reduction and guaranteeing compactness of hash codes, thus resulting in deteriorated performance.

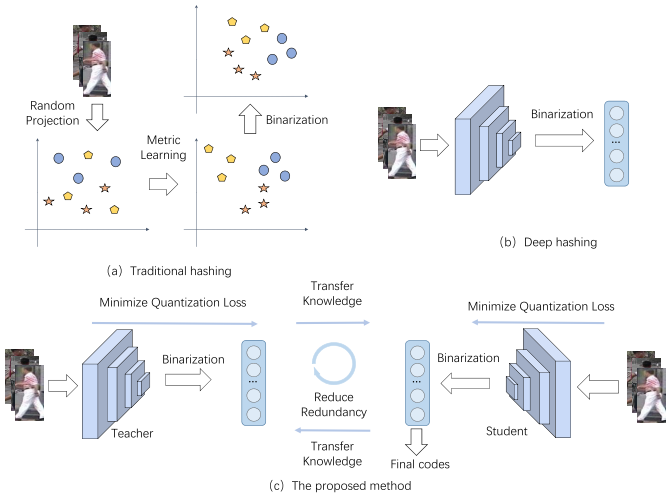


Fig. 1. Illustration of hashing ReID: (a) Traditional methods, (b) Deep hashing methods, (c) The proposed method.

Recently, online knowledge distillation [28], [29] has been widely used for model acceleration. Inspired by the recent advance in online knowledge distillation, we make an attempt to integrate mutual learning into the proposed framework. However, common mutual learning has two crucial drawbacks which limit its application. Firstly, it does more harm than good to the performance of large models by imposing strong constraints for minimizing the output difference between two models. Besides, the complementary effect of two models cannot be fully utilized by treating two models equally. Therefore, to tackle these challenges, we design a salience-guided iterative asymmetric mutual hashing network. As shown in Fig.1, the proposed method optimizes hash codes from two distinct views: a) reducing information redundancy across models, b) minimizing quantization loss within models. Specifically, SIAMH enables the network to generate hash codes based on the most informational regions, thus explicitly reducing the information redundancy and guaranteeing the compactness of hash codes. Besides, it can alleviate the drawbacks brought by mutual learning through asymmetric training, in which the student model can learn discriminative dark knowledge and meanwhile impose implicit regularization on the teacher model, resulting in complementary effects. Additionally, a diverse partition module and a self-distillation quantization loss are employed to extract effective real-valued features and close the discrepancy of similarity structure in the quantization process for a single model respectively. Finally, compared with existing methods, SIAMH can largely slow down the speed of accuracy decrease with code length becoming shorter by explicitly reducing the redundancy, as shown in Fig.2. The main contributions are summarized as follows:

1) We present a new salience-guided iterative asymmetric hashing framework, called SIAMH, in which a novel salience-guided self-distillation branch (SSB) is designed to make SIAMH generate binary codes from the continuously refined salient regions. By explicitly reducing the information redundancy, SSB can bring significant improvements for hash codes, especially for short ones.

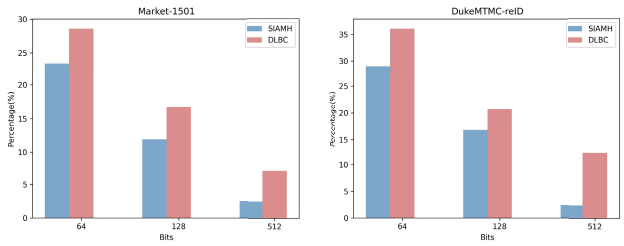


Fig. 2. Compared with 2048-bits results, accuracy decrease percentage of different bit length for SIAMH and DLBC [10].

2) A novel iterative asymmetric mutual training strategy, called IAMT, is proposed in which two models learn regularized dark knowledge and provide complementary effects for each other through asymmetric training to alleviate the drawbacks of mutual learning, thus further boosting final performance.

3) Extensive experiments on five widely used datasets show that the proposed method achieves superior performance to existing hashing ReID methods. For the state-of-the-art real-valued methods, SIAMH further closes the gap with them in terms of accuracy while significantly improves the retrieval efficiency.

The remainder of the paper is organized as follows: In Section II, we review some related works about knowledge distillation and hashing ReID and analyze their drawbacks. Section III elaborates on the proposed salience-guided iterative asymmetric mutual hashing network. Section IV presents the experimental results of the comparisons and evaluations. Finally, the conclusions are drawn in Section V.

II. Related Work

Since the proposed method mainly focuses on the design of hash based ReID networks and knowledge distillation, we briefly review related works in this section.

A. Traditional Person Re-Identification

Traditional person ReID methods [1]–[8], [30], [31] mainly address the pedestrian matching problems from two perspectives: 1) learning robust and effective view-invariant representations, which focus on dealing with cross-view appearance changes brought by various factors such as background clutter, occlusions, 2) constructing sophisticated and powerful feature extraction networks. Among them, Yu *et al.* [1] propose a hard-aware point-to-set loss to solve problems brought by traditional sampling. Bai *et al.* [2] develop a manifold-preserving algorithm by integrating manifold-based affinity learning. Li *et al.* [3] focus on integrating multiple attentions from diverse levels and optimizing multi-scale attention features selectiveness to enrich final results. In [4], a part-based baseline is proposed to maintain part-level content consistency for accurate part location. Bryan *et al.* [5] attempt to model long-range region relation by utilizing second-order statistics within features. Fu *et al.* [6] propose a horizontal pyramid matching strategy to alleviate influence brought by crucial

body parts missing problems. Zhang *et al.* [7] design a sophisticated alignment approach and jointly optimize global and local features in the proposed framework. Lin *et al.* propose a correspondence structure to address spatial misalignments brought by camera-view changes. More recently, [8] design an attentive but diverse network, aiming to fully promote the complementary powers of attention and feature diversity. Nevertheless, to achieve high matching accuracy, all of these works need complex frameworks to obtain high-dimensional real-valued features, resulting in a slow retrieval process. They are not scalable for a large number of gallery images. Consequently, it grows rapidly for the response time and memory costs with the size of the gallery set becoming larger.

B. Hash Based Person Re-Identification

Previous hash based ReID methods can be roughly divided into two categories: traditional methods [23], [27], deep hashing methods [10], [24]–[26], [32]–[34]. 1) As shown in Fig.1 (a), traditional methods aim to learn subspace transformation and binary coding schemes, which utilize multiple projection matrices to map high-dimensional real-valued features into Hamming space. These methods generally take each camera view as one modality and exploit the correlations between distinct camera sources to produce view-invariant representations. Among them, cross-view binary identities (CBI) [27] minimizes the intra-person Hamming distance and maximizes the cross-covariance to construct two sets of hash functions, while cross-camera semantic binary method (CSBT) [23] tries to alleviate the intrinsic cross-view variations. However, the objectives of these methods generally require complex non-convex optimization and need to explicitly design the complicated loss functions, resulting in unstable training processes and poor performance. 2) The basic framework of deep hashing methods is shown in Fig.1 (b). These approaches generally insert hashing layers at the end of networks to generate approximate binary codes. Specifically, hashing layers are mostly fully connected layers with tan-like activation. Among them, Bit-scalable deep hashing (DSRCH) [32] proposes a novel formulation of relative similarity comparison based on the triplet-based model, pursuing generating bit-scalable results. PDH [24] integrates a part-based model into deep hashing frameworks to improve local discrimination, while in [25], ABC designs an adversarial learning based implicit binary transformation strategy. Specifically, it uses Bernoulli distribution to implicitly guide the network to produce samples conforming to the same distribution. By explicitly encoding local visual entails, DLBC [10] formulates a binary-local semantic mutual information maximization term to close the gap with real-valued methods. CPDH [33] focuses on preserving consistency between hash codes and optimizes high-dimensional features in more robust directions. To accelerate the matching process, a coarse-to-fine search strategy is proposed by CtF [26] to enhance the quality of short codes.

Nevertheless, none of them focus on reducing the redundancy and guaranteeing compactness of hash codes, which is rather crucial for binary results, especially for short codes. Besides, they generally adopt simple networks for fast feature extraction. However, it limits the performance of real-valued features, leading to deteriorated hash codes. As a result, there exists a large gap in matching accuracy between hash based ReID methods and real-valued approaches. To address these problems, differing from existing frameworks, our work simultaneously optimizes the information redundancy and improves the discrimination of real-valued features. Concretely, the proposed salience-guided self-distillation branch and flexible iterative asymmetric mutual training strategy work collaboratively to further close the gap with real-valued methods in accuracy.

C. Knowledge Distillation

Knowledge distillation refers to the work of transferring knowledge from one (teacher) to another (student), which was first proposed by Hinton *et al.* [35]. It has been widely used in model compression and unsupervised learning in recent years. Existing methods of knowledge distillation can be roughly divided into three categories: online distillation, self-distillation, offline distillation. In online distillation, the student model and the teacher model are updated simultaneously and trained in an end-to-end mode. In the last two years, several online knowledge methods [29], [36] have been proposed. In [36], Zhang *et al.* propose to make a set of neural networks work collaboratively. As for the self-distillation [37], the teacher and student models share the same network. By utilizing the "soft targets" extracted from teacher models, online distillation can obtain latent "dark knowledge" to enhance the performance of the student model. The "soft targets" are often in the form of prediction results or intermediate features and contain pivotal similarity information for guiding the training process of the student model during distillation.

Recent progresses of online distillation have demonstrated its superiority over offline ones. Inspired by these methods, we make an attempt to integrate mutual learning strategy into the proposed framework. However, directly combining mutual learning may result in sub-optimal results. Two main factors can be obtained through analysis: 1) Firstly, due to the discrepancy in learning ability of two models, equally imposing constraints for closing their gap will harm the performance of the large model, thus producing deteriorated results. 2) Besides, common mutual learning cannot fully use the different advantages of two models. For example, large models perform well on training sets and are much more discriminative, while light models perform well on generalization tasks. Therefore, simply combining mutual learning will treat two models equally, thus reducing complementary effects. To address these issues, we propose a salience-guided iterative asymmetric mutual training strategy, which can fully enhance the performance of two models through iterative interaction.

III. Salience-Guided Iterative Asymmetric Hashing

A. Framework Overview

In this section, the framework of the proposed SIAMH method will be described in detail. For convenience, some frequently used notations are introduced firstly. Specifically, we

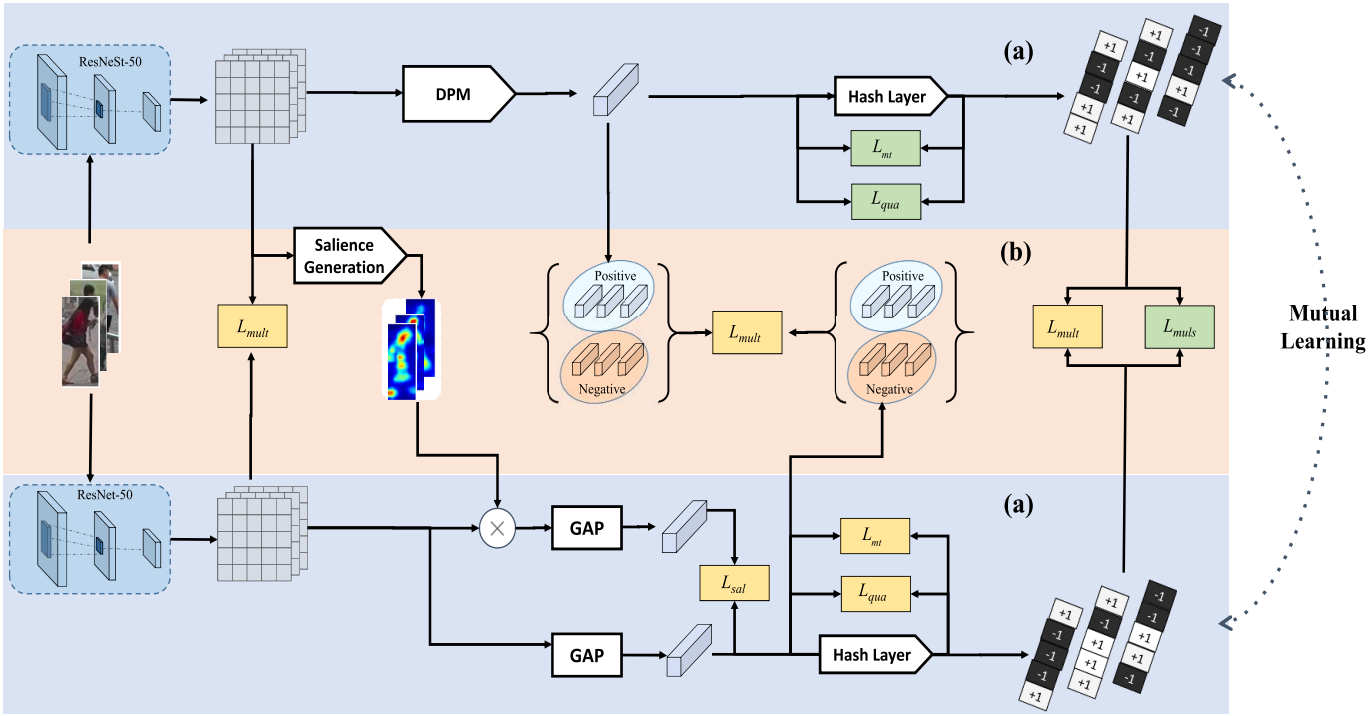


Fig. 3. The overall framework of the proposed SIAMH: (a) Self-distillation quantization loss minimization within models, (b) Salience-guided iterative asymmetric mutual training. DPM indicates the diverse partition module. Loss boxes with light yellow color are loss functions for the student model, while green for the teacher model.

denote global average pooling and global max pooling as GAP and GMP respectively. Fully connected layers and convolution layers are simplified as Fc and Conv. Besides, the whole training set is indicated as $\{P_i, y_i\}_{i=1}^{N}$, where P_i denotes the i-th image, $y_i \in \{1, 2, 3, \ldots, N_d\}$ represents the corresponding person identity, N and N_d indicate the image number and the number of pedestrian identities in the training set respectively. Besides, H are the real-valued features and B indicate hash codes. C is the batch size. I_t^H and I_s^H denote the prediction results of real-valued features from the teacher model and the student model respectively, while I_t^B and I_s^B indicate the corresponding classification results of binary codes. Subscript s and t represent the result is from the student model and the teacher model respectively. $sgn(\cdot)$ is an element-wise sign function, which outputs -1 for positive values and 1 otherwise. $tanh(\cdot)$ is calculated as follows:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

As shown in Fig.3, the proposed SIAMH adopts the general pipeline of deep hashing methods within a single model. Concretely speaking, a widely-used convolution network is utilized as the backbone to generate feature maps. Specifically, ReNeSt-50 and ResNet-50 are exploited here for the teacher and student model respectively. Afterward, a pooling layer is followed to produce high-quality but redundant real-valued features. Finally, it is successively inputted into Fc layer for dimension consistency with binary codes and then transformed into discrete hash codes $B_i \in \{-1, 1\}^q$ via hash layers $\mathbb{H}(\cdot)$, where q indicates the length of hash codes. Specifically, we denote the real-valued features before hash layers as $H_i \in \mathbb{R}^q$. Owing to the ill-gradient problem of $sgn(\cdot)$, hash layers are generally Fc layers with tanh-like activation, resulting in sub-optimal approximate discrete codes and an extra quantization procedure in the testing stage. Unlike them, the above hash layers are replaced with greedy hash layers in SIAMH to directly generate completely discrete results.

Different from existing hashing ReID methods, SIAMH performs binary code optimization from two distinct perspectives: quantization loss minimization within single models, redundancy reduction across models. For the first view, a diverse partition module and a self-distillation quantization loss are employed. The former focuses on generating highly approximate discrete real-valued features. Besides, it can explicitly establish the connections between final hash codes and multi-granularity visual details. The self-distillation quantization loss aims at closing the gap in similarity structure between hash codes and real-valued features. The above components work collaboratively for quantization loss minimization within single models and enhancing the quality of codes as a result. For the second view, a salience-guided iterative asymmetric training strategy is designed. Firstly, a salience-guided self-distillation branch (SSB) is proposed to maintain salience region consistency in the mutual learning process, which enables the student model to generate salience-based features and reduce information redundancy as a result. Additionally, the training of SIAMH is conducted in a novel iterative asymmetric mutual training mode, which can alleviate drawbacks of common mutual learning. Specifically, two models learn collaboratively and transfer regularized dark knowledge for

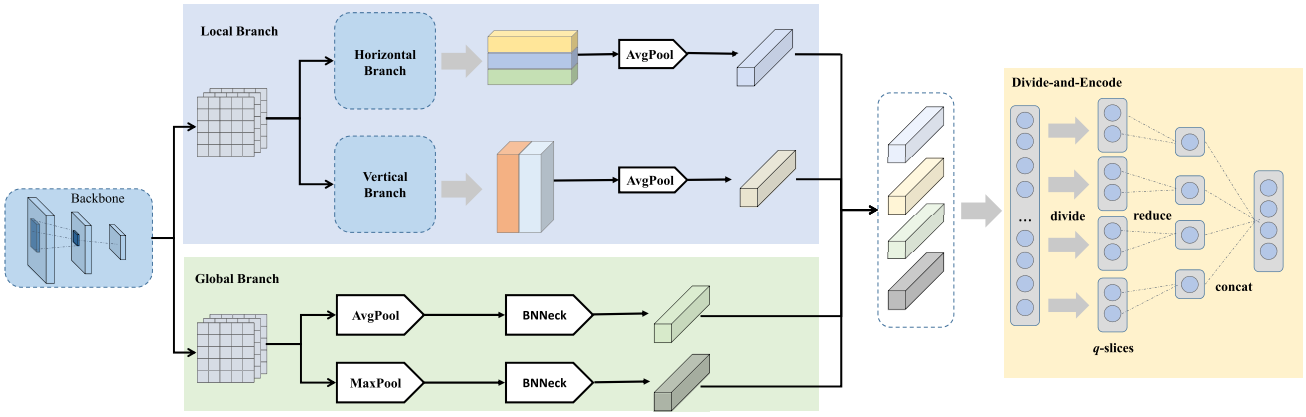


Fig. 4. The overall framework of the DPM.

each other throughout the training process. For the following content of this section, we will mainly describe the two above procedures of SIAMH, together with the optimization of the whole framework.

B. Self-Distillation Quantization Loss Minimization

1) Diverse Partition Module: Previous work [10], [38] has proved that high-quality real-valued features lead to smaller quantization loss. Therefore, to address this issue, a diverse partition module (DPM) is designed in the teacher model, which can improve features diversity and alleviate the influence of misalignment and occlusion meanwhile. As shown in Fig.4, the DPM consists of two branches: the global branch, the part branch in our implementation. In the global branch, two different pooling strategies are adopted to obtain global representations. Firstly, the initial $2048 \times 24 \times 8$-tensor is transformed into two 2048-dimension vectors via GAP and GMP respectively. Afterward, to optimize the embeddings from different spaces, a BNNeck [39] is employed for two branches, which consists of a classification layer and a batch normalization layer. Results of GAP and GMP strategies are indicated as $f_{gap}/f_{gmp} \in \mathbb{R}^{2048}$ respectively. For the part branch, it is composed of two branches as well. Firstly, the initial $2048 \times 24 \times 8$-tensor is split into three parts horizontally and vertically, which focuses on learning sub-features of different regions and explicitly encoding local visual clues. Therefore, 5 2048-dimension sub-features are obtained totally. Finally, these sub-features are concatenated into a single vector as final outputs of local branches, denoted as f_p. These global and local features are then inputted into the divide-and-encode module, which aims at guaranteeing the bit-independence of hash codes, thus reducing the quantization loss. Specifically, the feature is divided into q slices with equal length, where q denotes the length of hash codes. Then a fully-connected layer is used to map each slice into a one-dimension vector. Afterward, these one-dimension vectors are concatenated as the final output of the diverse partition module.

2) Self-Distillation Quantization: Existing methods mainly impose L2-norm constraints to close the gap between real-valued features and hash codes. However, since the discrepancy between them in information capacity, such hard explicit principles may inevitably harm the performance of real-valued features. Instead, we choose to mimic the difference in prediction distribution. The prediction scores of real-valued features are used as a form of softened class scores to guide the self-distillation process, where the former is detached from training when fed into loss functions. Specifically, the classical Kullback-Leibler divergence loss is utilized for it, which is shown below:

$$L_{kl}(m_1, m_2) = \sum_{i=1}^{N_d} m_1(c_i) \log\left(\frac{m_1(c_i)}{m_2(c_i)}\right) \tag{2}$$

where m_1, m_2 denote prediction results from the student and teacher models respectively, c_i is the i-th class. Moreover, the greedy hash layer is utilized to perform discrete optimization and directly generate binary codes. Specifically, $sgn(\cdot)$ is used as the activation function of hash layers. By minimizing loss function as Eq.(3), the framework can foresee the error in testing and propagate gradients of hash codes to the former layer. As a result, real-valued features and binary codes can be optimized in similar paces, thus further closing their gap.

$$L_{gr} = ||H - B||_p^p \tag{3}$$

where p is p-norm. Therefore the self-distillation quantization loss can be found in Eq.(4).

$$L_{qua} = L_{gr} + L_{kl} \tag{4}$$

C. Salience-Guided Iterative Asymmetric Mutual Training

The salience-guided iterative asymmetric mutual training (SIAMT) framework aims to simultaneously enhance the discrimination of hash codes and reduce information redundancy across models, as shown in Fig.3. Specifically, it consists of two crucial components: salience-guided distillation branch and iterative asymmetric mutual training strategy. The former focuses on enabling the light-weight student model to generate codes based on the salient regions, thus reducing redundancy between codes, while the latter performs iterative asymmetric mutual training, in which two models provide different outputs as dark knowledge for each other to alleviate the drawbacks of common mutual learning and fully utilize the complementary advantages of two models meanwhile.

1) Salience-Guided Self-Distillation Branch: Most hashing methods generally ignore the information redundancy between hash codes and perform feature extraction based on the whole image. As a result, visual clues that are irrelevant to final retrieval $e.g.$ the clutter background are also included between hash codes, resulting in severe redundancy. Due to the weak information capacity of binary codes, such feature extraction procedures may harm the performance, especially for short codes. To address this issue, a salience-guided self-distillation branch (SSB) is proposed in our model. Specifically, feature maps from two branches of the teacher model are added and averaged firstly. Afterward, the results are fed into the salience generation block to produce class activation maps. Subsequently, the obtained maps are transformed into probability maps through Softmax function. The values in the maps are set to zero when they are lower than the predefined threshold, otherwise remain the same. Then these weighed maps are resized and multiplied with feature maps of the student model to filter out background disturbances. The above process can be summarized as follows:

$$F_{sal} = Re\left(In\left(L_{CAM}\right)\right) \odot F \tag{5}$$

where $F \in \mathbb{R}^{C\times H\times W}$ is the original feature maps of channel number C, height H and width W. $\odot$ is the dot-wise multiply operation. $L_{CAM} \in \mathbb{R}^{3\times H_I\times W_I}$ denotes the activation maps, H_I, W_I indicate the height and width of input images. $In(\cdot)$ and $Re(\cdot)$ are the interpolation and reshape operations, which are used to convert the activation maps to the same size as feature maps. F_{sal} denotes the transformed results. A GAP layer is followed to flatten the former result and obtain single vectors. Finally, the classification scores of SSB are utilized as pseudo labels to explicitly guide the training of the GAP branch, which is used for the final generation of hash codes. The corresponding loss is shown as below:

$$L_{sal} = L_{kl}(I_{gap}, I_{sal}) \tag{6}$$

where I_{gap}, I_{sal} are the prediction results of the GAP and salient branches respectively. In this way, the GAP branch can alleviate the negative effects of uninformative issues and generate hash codes based on the salient regions, thus explicitly reducing the redundancy. Finally, in the prediction stage, the student model does not rely on activation maps from the teacher model. Therefore, it brings no extra time costs.

2) Iterative Asymmetric Mutual Training: On the one hand, owing to the constraint of fast feature extraction, the structure of student model should be lightweight. However, it limits further boost in performance within a single model. On the other hand, due to the gap in learning ability between two models, directly integrating mutual learning cannot fully exploit their complementary effects and may even harm their performance. Therefore, to tackle these issues, an iterative asymmetric mutual training strategy is proposed. Firstly, to fully utilize the powerful learning ability of the teacher model, the student model is trained from multiple views based on multi-level outputs of the teacher model, including feature maps, real-valued features, hash codes. The first view mainly focuses on minimizing the discrepancy in activation patterns for the teacher and student models when receiving similar inputs. To be specific, directly optimizing the Euclidean distance between their feature maps may bring large model bias, thus resulting in overfitting to outputs of the teacher model. Instead, the pairwise activation similarities of input samples between them are adopted to be mimicked, which are shown as below:

$$L_{map}(F_s, F_t) = \left\| \frac{F_s \cdot {F_s}^T}{\left\| F_s \cdot {F_s}^T \right\|_2} - \frac{F_t \cdot {F_t}^T}{\left\| F_t \cdot {F_t}^T \right\|_2} \right\|_{\mathbb{F}}^2 \tag{7}$$

where F_s, F_t are feature maps from the student model and the teacher model. $\mathbb{F}$ denotes the Frobenius-norm. Concerning the real-valued features level, it is optimized from three perspectives: classification score, pairwise and cosine similarities. For the former, a similar function as Eq.(2) is used. The cosine similarities between real-valued features from two models are optimized as well, whose formulation is in Eq.(8).

$$L_{cos}(H_s, H_t) = \eta\left(1 - \frac{H_s \cdot H_t}{\| H_s \|_2 \| H_t \|_2}\right) \tag{8}$$

where η is a weight parameter. Finally, the pairwise similarities are also included as below:

$$L_{sim} = \sum_{i,j} \frac{1}{q} \left\| E_t^{i,j} - \widehat{E}_s^{i,j} \right\|^2 \tag{9}$$

where $E_t^{i,j}/\widehat{E}_s^{i,j}$ is the Euclidean distance between H_t^i and H_s^j, H_t^i/H_s^i is the feature of image x_i from the teacher/student model. q denotes the feature length. For real-valued feature level, the losses can be summarized as follows:

$$L_{min} = L_{kl} + L_{cos} + L_{sim} \tag{10}$$

As for hash codes, a similar operation to real-valued feature level is adopted. Therefore, to be summarized, the objective function of student model in the IAMT is shown as follows:

$$L_{mult} = L_{min}(H_s, H_t) + L_{min}(B_s, B_t) + L_{map}(F_s, F_t) \tag{11}$$

Additionally, to fully utilize the generalization of lightweight student models, the classification scores I_t^H, I_t^B from the student model are used as dark knowledge to impose regularization on the teacher model, which can maintain its discrimination meanwhile. Therefore, the objective function of teacher model in the IAMT is shown as follows:

$$L_{muls} = L_{kl}(I_t^H, I_s^H) + L_{kl}(I_t^B, I_s^B) \tag{12}$$

Finally, as shown in Fig.5, the weight maps from SSB can boost the performance of hash codes from the student model, thus enabling it to generate better classification results. Meanwhile, these results will impose stronger regularization effects on the teacher model. With the gradient backpropagation, more discriminative feature maps can be obtained, thus making the SSB generate more refined weighted maps. Therefore, in this way, SSB can be trained in a positive loop iteration mode, in which the salient regions can be continuously mined and refined to further improve the discrimination of final results.

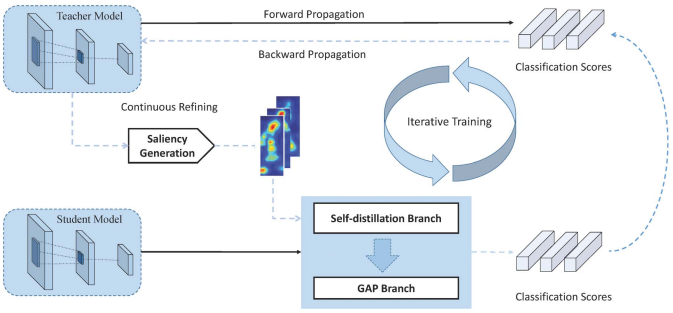


Fig. 5. The proposed iterative asymmetric training strategy.

D. Optimization

1) Overall Objective Function: Inspired by the recent progresses of classification and triplet losses on ReID, they are included in our objective function. Specifically, the classification loss is calculated as follows:

$$L_{sf} = -\frac{1}{C}\sum_{i=1}^{C} y_i \log \frac{e^{W_{Ii}^T f_i + b_{Ii}}}{\sum_{j=1}^{N_d} e^{W_{Ij}^T f_i + b_{Ij}}} \tag{13}$$

where W, b are the weights and bias for classification layers. The formulation of triplet losses can be found in Eq.(14).

$$L_{tri} = \sum_{i=1}^{C} [m + D(f_\theta(I^{i_a}), f_\theta(I^{i_p})) - D(f_\theta(I^{i_a}), f_\theta(I^{i_n}))]_+ \tag{14}$$

where m is the margin parameter. $f_\theta(\cdot)$ denotes the feature extraction process. I^{i_a} is the $i-th$ image in the batch, I^{i_p} and I^{i_n} indicate the hardest positive and negative image of I^{i_a}. $D(\cdot)$ is the Euclidean distance function. For convenience, the sum of the above loss for H and B is denoted as follows:

$$L_{mt} = L_{sf}(I^B, y) + L_{sf}(I^H, y) + L_{tri}(H, y) + L_{tri}(B, y) \tag{15}$$

Therefore, the overall objective function for the teacher model can be summarized as follows:

$$L_{th} = L_{mt} + L_{muls} + L_{qua} \tag{16}$$

The loss function for the student model is listed as below:

$$L_{stu} = L_{mt} + L_{mult} + L_{qua} + L_{sal} \tag{17}$$

2) Training and Testing: The whole training process is summarized as Algorithm.1. The whole framework is trained in an iterative asymmetric mutual learning mode. In the prediction stage, final hash codes for out-of-sample images can be directly obtained through the GAP branch of the lightweight student model, while having no extra binarization operation. Finally, the target images are retrieved using the Hamming distance metric.

Algorithm 1 Learning Algorithm for SIAMH

Input: $P = \{p_i\}_{i=1}^N$: images for training; y: person identities for training images; student model: S; teacher model: T; learning rate for S: μ_s; learning rate for T: μ_t; random sample: R; batch size: M; number of iteration k←0; maximum iteration number: K;

Output: Parameters of S: θ_s

Initialize parameters of S and T: θ_s and θ_t with pre-trained parameters on ImageNet.
repeat
  $\{p^i\}_{i=1}^M \leftarrow R(P)$
  $B_t, H_t, F_t, I_t^B, I_t^H \leftarrow T(\{p^i\}_{i=1}^M)$
  $F_{sal} \leftarrow Re\,(In\,(L_{CAM})) \odot F$
  $B_s, H_s, F_s, I_s^B, I_s^H \leftarrow S(F_{sal}, \{p^i\}_{i=1}^M)$
  Compute L_{th} by Eq.16.
  Compute L_{stu} by Eq.17.
  Update $\theta_s \leftarrow \theta_s - \mu_s \frac{\partial}{\partial \theta_s} L_{stu}$
  Update $\theta_t \leftarrow \theta_t - \mu_t \frac{\partial}{\partial \theta_t} L_{th}$
until $k > K$
return θ_s

IV. Experiments

In this section, we first describe datasets and evaluation protocols. Specifically, for fair comparison in retrieval time, we introduce the hardware platform that our method is implemented on. Then we evaluate the proposed SIAMH and conduct comparisons with state-of-the-art real-valued and hashing person ReID methods to demonstrate the effectiveness and efficiency of SIAMH.

A. Datasets and Evaluation Measures

We conduct experiments on five off-the-shelf datasets Market1501 [49], DukeMTMC-ReID [50], CUHK03 [51], MSMT [52], and LaST [53]. The details are shown as follows.

Market1501 includes 32,688 bounding boxes of 1,501 pedestrians captured from 2-6 cameras. We split 12,936 images of 751 persons for training, and use the 3,368 images for the query. The left 19,732 images are utilized for the gallery.

DukeMTMC-ReID contains 36,411 images from 1,812 persons under 8 cameras. It provides a fixed training/split strategy with 16,522 images of 702 persons for training, 2,288 images for the query, and the left 17,611 images for the gallery.

CUHK03 consists of 14,096 images of 1,467 identities captured by 6 surveillance cameras. For the dataset, we follow the widely utilized protocol proposed in [50] with the 20 training/testing splits.

MSMT17 contains 126,441 images of 4,101 identities. All the images are captured from 15 cameras. Following the standard protocol [52], 1,041 identities are split for training and the rest for test.

LaST contains 228,156 images of 10,862 identities. 70,923 images of 5,000 identities are split as the training set, while 20,584 images of 56 identities are used as the evaluation set. The rest 133,214 images from 5,803 identities are adopted as the test set.

Evaluation Metrics: Following the standard evaluation protocol in ReID, we use the Cumulated Matching

TABLE I

COMPARISON WITH STATE-OF-THE-ART REAL-VALUED METHODS ON MARKET-1501, DUKEMTMC-REID, AND MSMT17. B AND R DENOTE BINARY AND REAL-VALUED FEATURES RESPECTIVELY. BEST RESULT IS SHOWN IN BOLD FORMAT. Q.TIME DENOTES TIME COSTS FOR QUERY

Methods	Code		Market-1501			DukeMTMC-reID			MSMT17		
	Type	Length	R1 (%)	mAP (%)	Q.Time (s)	R1 (%)	mAP (%)	Q.Time (s)	R1 (%)	mAP (%)	Q.Time (s)
PN-GAN[40]	R	1024	89.4	72.6	-	73.6	53.2	-	-	-	-
IDE[41]	R	2048	88.1	72.8	-	69.4	55.4	-	-	-	-
BoT[39]	R	2048	94.1	85.7	2.2	86.4	76.4	2.0	-	-	-
SPReID[42]	R	10240	92.5	81.3	-	84.4	71	-	-	-	-
PCB[4]	R	12288	93.8	81.6	6.9	83.3	69.2	6.3	68.2	40.4	1.4×10^{2}
VPM[43]	R	14336	93	80.8	-	83.6	72.6	-	-	-	-
ABD-Net[8]	R	3072	95.6	88.3	2.8	89.0	78.6	2.5	82.3	60.8	7.7×10^{1}
RGA-SC[44]	R	2048	96.1	88.4	2.4	-	-	-	80.3	57.5	6.3×10^{1}
SCSN[45]	R	1536	95.7	88.5	-	**90.1**	79.0	-	83.0	58.0	-
SONA[5]	R	1536	**95.7**	**88.8**	-	89.6	78.3	-	-	-	-
Osnet[46]	R	512	94.8	84.9	1.8	86.6	73.5	1.5	78.7	52.9	4.3×10^{1}
DGNet[47]	R	1024	94.8	86.0	-	86.6	74.8	-	77.2	52.3	-
IANet[48]	R	2048	94.4	83.1	-	87.1	73.4	-	75.5	46.8	-
SIAMH	B	2048	95.4	**88.8**	$\mathbf{2.8\times10^{-1}}$	**90.1**	**79.4**	$\mathbf{2.0\times10^{-1}}$	**83.2**	**62.5**	$\mathbf{6.8\times10^{-1}}$

Characteristics (CMC) curve and the mean average precision (mAP) to evaluate the performance. The single query setting is applied in all experiments.

B. Implementation Details

We conduct experiments based on Fastreid [54], a popular framework for deep-learning ReID in Pytorch. The standard Adam [55] algorithm is adopted. Initial learning rate is set to 3.5×10^{-4} and decaying to 0.1 at 20 and 40 epochs. For a fair comparison, the backbone networks for all the compared methods and experiments are ResNet-50 and pre-trained on the ImageNet dataset. The input images are resized to 384×128. The batch size in the training stage is 64 and we set the batch size for testing to 128. The training epoch number is set to 120. The images for training are augmented by random erasing and random flipping. Besides, the XBM [56] training strategy is adopted. The hardware platform for our method and all the compared methods is a PC with Intel Core i5 CPUs (2.6GHz), and three NVIDIA GTX 2080Ti GPUs with 11G memory.

In our evaluations, we use the Local Maximal Occurrence (LOMO) feature [57] and deep features extracted by ResNet-50 to evaluate the non-deep hashing methods. Following the standard protocol, we randomly choose 1,000 samples from the training set as the anchor images in KSH [58], SDH [12] to construct kernels. For computational efficiency, as for the LOMO feature, we use PCA to reduce the feature dimension to 3,000. For all the hashing and real-valued methods, we follow the suggestion of the corresponding authors to set the related hyper-parameters. The source code of KSH, SDH is not available, so we carefully re-implement them according to the setting of their authors. For SIAMH, we set β=0.1 for both teacher and student models. We set λ=1.5, η=0.3 to make the distillation loss contribute more to the total loss.

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON LAST. B AND r DENOTE BINARY AND REAL-VALUED FEATURES RESPECTIVELY. BEST RESULT IS SHOWN IN BOLD FORMAT

Methods	LaST				
	Type	R1 (%)	R5 (%)	R10 (%)	mAP (%)
PCB[4]	R	50.6	68.0	73.9	15.2
Osnet[46]	R	64.3	78.9	82.6	21.0
ABD-Net[8]	R	48.5	67.6	74.4	16.1
CtF[26]	B	70.0	83.3	86.7	26.5
SIAMH	B	**72.1**	**85.1**	**88.3**	**30.2**

C. Comparisons With Real-Valued Methods

As shown in Table I and II, we compare our method with state-of-the-art real-valued ReID methods, which generally adopt high-dimensional real-valued features for better performance. The results show that SIAMH outperforms the majority of real-valued methods and meanwhile largely improves the matching efficiency. Specifically, compared with the popular method BoT, SIAMH achieves 3.05%, 3.01% improvement in Rank-1 on two datasets. Though SONA and ABD-Net produce slightly better results than SIAMH, it further closes the gap between real-valued methods and binary methods and even achieves better results than most real-valued methods in some cases. Additionally, it can be discovered that methods with longer features generally show better performance but suffer from low matching efficiency. For example, as for methods adopting features shorter than 2048-dimension, they all achieve Rank-1 scores less than 90% and 80% on Market1501 and DuekMTMC-reID respectively. However, methods using features longer than 10,240-dimension generally

TABLE III

COMPARISON WITH TRADITIONAL HASH BASED METHODS IN map(%). "sdh"/"sdh+cnn" DENOTES sdh USING HAND-CRAFT AND DEEP FEATURES RESPECTIVELY. BEST PERFORMANCE IS SHOWN IN BOLDFACE

Methods	Market1501				DukeMTMC-ReID				CUHK03			
	32bits	64bits	96bits	128bits	32bits	64bits	96bits	128bits	32bits	64bits	96bits	128bits
COSDISH[11]	1.89	3.68	4.83	5.94	1.02	2.39	3.81	5.11	0.82	1.54	2.59	3.01
SDH[12]	1.65	2.93	3.78	4.06	0.98	1.89	2.25	2.42	1.00	1.24	1.32	1.65
KSH[58]	4.66	5.62	6.16	6.20	2.13	2.67	3.31	3.34	2.86	2.53	2.11	1.75
ITQ[59]	1.70	3.00	3.83	4.43	0.91	1.41	1.77	2.16	0.68	0.76	0.82	0.95
LSH[13]	0.44	0.83	1.18	1.68	0.40	0.58	0.83	1.06	0.37	0.46	0.44	0.68
COSDISH+CNN	0.79	1.06	1.47	1.82	0.62	1.09	1.42	1.79	0.39	0.57	0.63	0.62
SDH+CNN	0.73	1.26	1.55	1.67	0.66	0.89	1.06	1.40	0.44	0.65	0.63	0.63
KSH+CNN	0.77	0.74	0.54	0.68	0.30	0.37	0.44	0.46	0.49	0.41	0.33	0.41
ITQ+CNN	0.77	1.07	1.21	1.29	0.56	0.92	1.21	1.34	0.38	0.38	0.43	0.45
LSH+CNN	0.50	0.77	1.04	1.27	0.48	0.74	0.99	1.17	0.33	0.35	0.35	0.42
SIAMH	**45.6**	**65.7**	**73.9**	**78.4**	**33.4**	**53.30**	**61.9**	**64.6**	**36.1**	**54.3**	**61.8**	**67.4**

achieve higher than 92% results in Rank-1. Besides, methods using longer features generally need much longer time for a single query. Take PCB as an example, it adopts 12,288-dimension features and needs 6.9s and 6.3s for a single query on two datasets respectively. In contrast, BoT only takes around 2.0 seconds. Finally, it can be revealed from the table that binary codes of SIAMH can achieve nearly the same performance with state-of-the-art real-valued methods but bring a large reduction to query times costs. When applied in larger gallery set such as MSMT17 and LaST, SIAMH shows more outstanding advantages both in accuracy. Specifically, when compared with the ABD-Net, the proposed SIAMH outperforms it by 0.9% and 1.7% in Rank-1 and mAP on MSMT17 respectively.

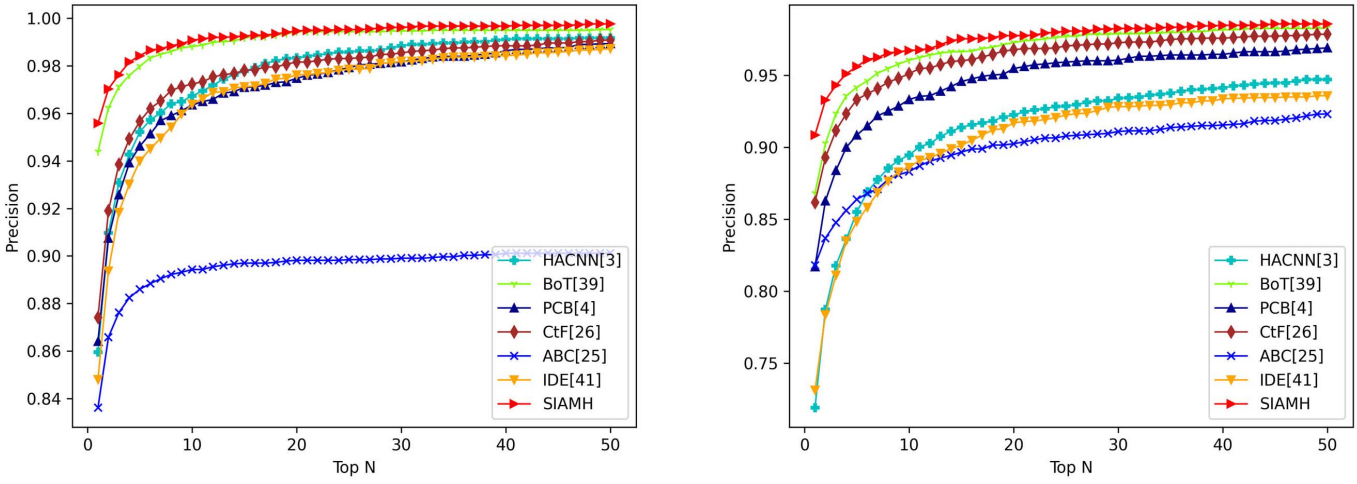
D. Comparisons With Hashing Methods

In this section, we compare SIAMH with other state-of-the-art hashing methods including traditional and deep hash approaches, which are shown in Table III and IV. Because all methods have the same query time when the code length is the same, here we only compare time costs for feature extraction. As the results reveal, traditional hashing methods generally perform much more poorly than deep hashing ones. For example, KSH and ITQ perform best among traditional methods, while achieving half the worst result in deep hashing methods. Additionally, when using deep features, traditional features even perform worse. Two factors may account for it: 1) Firstly, deep features contain biased knowledge of deep networks and complex semantic structure, which is rather abstract for traditional methods. Therefore, discriminative information cannot be fully extracted from deep features for them. 2) Besides, training processes of traditional methods are hard to converge. By contrast, low-level visual clues are utilized to extract hand-crafted features, which can be easily repeated by calculating projection matrixes. As a result, compared with deep features, hand-craft ones are more suitable for traditional methods. Besides, owing to limited information capacity of binary codes, there still exists a gap between real-valued methods and hashing ones, especially for short codes. For example, when adopting 2048-dimension features, DLBC and CtF achieve relatively high performance. However, their accuracy drops rapidly and has a large gap with real-valued methods when using codes shorter than 512 bits, which is mainly due to the severe information redundancy between their codes. By contrast, SIAMH can achieve higher performance for long hash codes owing to its discriminative feature extraction. Additionally, by explicitly reducing the redundancy between hash codes, SIAMH can alleviate the performance deterioration and further close the gap in accuracy between these two types of methods, especially for short codes. Specifically, compared with the state-of-the-art hashing method DLBC, SIAMH outperforms it by 5.5% and 1.4% in mAP when using codes of 512 and 2048 dimensions. More importantly, with the code length reducing, the gap between SIAMH and DLBC becomes larger as well. When comes to more challenging datasets such as DukeMTMC and MSMT, SIAMH shows greater advantages than DLBC. As for CtF, though it can extract multiple hash codes in a single framework, SIAMH outperforms it both in accuracy and time costs for feature extraction. With regard of other hashing methods, SIAMH can achieve better performance than the majority of them even when they use longer codes. For example, ABC using 2048-dimension codes achieves 64.7% and 81.4% for Rank-1 and mAP on Market-1501 while SIAMH of 128 bits obtains 78.4% and 90.6% respectively, which boosts the former by 13.7% and 9.2%. Finally, compared with existing methods, SIAMH shows great model flexibility by integrating iterative asymmetric mutual learning, which will be verified in the following experiments. Apart from the mAP and Rank-1, we compare some of these methods with SIAMH in CMC-50 curve, which is a more intuitive form. As shown in Fig.6, SIAMH outperforms the binary and real-valued methods across all the coordinates, which further demonstrates the superiority of SIAMH.

TABLE IV

COMPARISON WITH STATE-OF-THE-ART DEEP HASHING METHODS ON MARKET-1501, DUKEMTMC-REID, AND msmt17 IN RANK-1(%), mAP(%) AND QUERY TIME(s). BEST RESULT IS SHOWN IN BOLDFACE. F.TIME DENOTES TIME COSTS FOR FEATURE EXTRACTION

Methods	Code	Market-1501			DukeMTMC-reID			MSMT17		
	Length	R1 (%)	mAP (%)	F.Time (s)	R1 (%)	mAP (%)	F.Time (s)	R1 (%)	mAP (%)	F.Time (s)
DRSCH [32]	512	17.1	11.5	-	19.3	13.6	-	-	-	-
DSRH [14]	512	27.1	17.7	-	25.6	18.6	-	-	-	-
HashNet [60]	512	29.2	19.1	-	40.8	28.6	-	-	-	-
DCH [15]	512	40.7	20.2	-	57.4	37.3	-	-	-	-
CSBT [23]	512	42.9	20.3	-	47.2	33.1	-	-	-	-
PDH [24]	512	44.6	24.3	-	-	-	-	-	-	-
DeepSSH [20]	512	46.5	24.1	-	-	-	-	-	-	-
ABC [25]	2048	81.4	64.7	**3.3×** 10^{-2}	82.5	61.2	**3.0×** 10^{-2}	-	-	-
CPDH [33]	2048	89.5	77.1	-	81.6	66.4	-	-	-	-
CtF [26]	32	60	37.7	3.9× 10^{-2}	49.5	28.7	3.7× 10^{-2}	33.0	14.7	4.0× 10^{-2}
	128	88.9	71	3.8× 10^{-2}	78.6	59.4	3.7× 10^{-2}	64.2	37.9	4.0× 10^{-2}
	512	92.8	82.2	3.8× 10^{-2}	85.4	71.6	3.8× 10^{-2}	73.2	49.0	4.0× 10^{-2}
	2048	93.7	85.4	3.9× 10^{-2}	87.7	75.7	3.7× 10^{-2}	76.7	52.5	4.0× 10^{-2}
DLBC [10]	64	82.7	62.5	-	72.4	50.6	-	-	-	-
	128	89.7	72.8	-	80.5	62.2	-	-	-	-
	512	92.1	81.2	-	84.5	68.8	-	-	-	-
	2048	94.6	87.4	-	88.7	78.5	-	78.2	55.6	-
SIAMH	32	65.6	45.6	3.2× 10^{-2}	53.9	33.4	3.0× 10^{-2}	33.4	15.1	3.5× 10^{-2}
	64	83.0	65.6	3.3× 10^{-2}	72.5	53.3	3.0× 10^{-2}	54.2	29.8	3.5× 10^{-2}
	128	90.6	78.4	3.3× 10^{-2}	81.2	64.6	3.0× 10^{-2}	69.2	42.5	3.5× 10^{-2}
	512	94.8	86.7	3.3× 10^{-2}	88.7	77.3	3.0× 10^{-2}	78.7	54.5	3.5× 10^{-2}
	2048	**95.4**	**88.8**	**3.3×** 10^{-2}	**90.1**	**79.4**	**3.0×** 10^{-2}	**83.2**	**62.5**	**3.5×** 10^{-2}



(a) CMC-50 on Market-1501

(b) CMC-50 on DukeMTMC-reID

Fig. 6. The CMC-50 Curve results on Market-1501 and DukeMTMC-reID.

Component Analysis: In this section, extensive ablation studies are performed for both teacher and student models to verify the effect of each component. Concretely, the network which only consists of ResNet-50 and a greedy hash layer is selected as the ReID baseline. Moreover, we conduct experiments to verify the effectiveness of self-distillation quantization loss (SD), salience-guided self-distillation branch (SSB), and the iterative asymmetric mutual training (IAMT).

TABLE V
ABLATION STUDIES OF SIAMH ON MARKET-1501 WITH 128 BITS

Methods	Market-1501			
	R1 (%)	R5 (%)	R10 (%)	mAP (%)
baseline	82.04	93.62	96.50	65.12
baseline+MT	77.38	90.94	94.24	59.30
baseline+SD	82.71	94.20	96.55	65.73
baseline+IAMT	85.60	95.75	97.18	74.73
baseline+DPM+IAMT	86.41	96.12	97.30	73.92
baseline+SSB	84.80	95.24	96.93	68.27
SIAMH	**90.59**	**96.70**	**98.51**	**78.43**

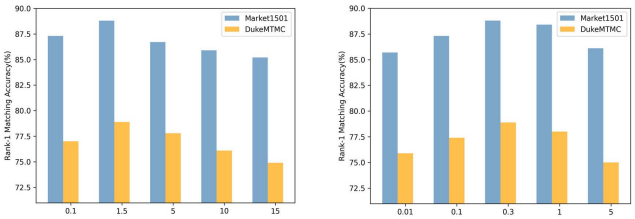
TABLE VI
ABLATION STUDIES OF THE TEACHER MODEL ON MARKET-1501 WITH 128 BITS. "TEACHER1-4" DENOTE THE VARIANTS OF THE TEACHER MODEL. "TEACHER" IS THE MODEL THAT USES ALL THE COMPONENTS

Methods	Market-1501			
	R1 (%)	R5 (%)	R10 (%)	mAP (%)
Teacher-1	88.09	95.58	97.36	73.12
Teacher-2	87.00	95.04	97.00	72.36
Teacher-3	86.70	95.31	97.18	72.55
Teacher-4	89.58	96.41	97.71	74.45
Teacher	**91.72**	**96.97**	**98.13**	**80.15**

Additionally, we add successive experiments, which focus on evaluating the impact of the diverse partition module to IAMT. Besides, we also list the performance of the common symmetric mutual learning strategy. Finally, we compare the results of all the variants with SIAMH. As summarized in Table V, symmetric mutual learning does harm to the performance of the baseline model owing to its drawbacks. As a comparison, the proposed IAMT can bring 3.6%/7.6% improvements to the baseline on Market-1501. The DPM can further boost the performance, which aims at encoding local visual clues to improve the discrimination of real-valued features. Additionally, SD and SSB can also enhance the performance of SIAMH by optimizing the quantization loss and explicitly reducing the information redundancy respectively. When seamlessly integrating these components, SIAMH can achieve its best performance. Besides the ablation study of SIAMH, we also conduct extensive experiments on the components of the teacher model to verify their effectiveness. As shown in Table. VI, Teacher-1 denotes the model without the local branch, while Teacher-2 is the model which removes the GAP branch from the original framework. Teacher-3 represents the model without the GMP branch and Teacher-4 is the variant without the Divide-and-Encode module. The results of Table. VI can effectively demonstrate the improvement of each component in the teacher model to the final performance. Finally, we conduct cross-domain experiments to verify the effect of IAMT on the teacher model, as shown in Table. VII. As the results show, by adopting the IAMT, the generalization capability of the teacher model is improved, which demonstrates the complementary effect of IAMT to both the teacher and student models.

TABLE VII
THE 128-BITS PERFORMANCE OF TEACHER MODEL ON CROSS-DOMAIN TASKS

Methods	DukeMTMC->Market1501			
	R1 (%)	R5 (%)	R10 (%)	mAP (%)
Teacher	31.38	52.05	60.78	13.90
Teacher+IAMT	**32.54**	**52.91**	**60.90**	**14.52**



(a) Impact of λ (b) Impact of η

Fig. 7. Impact to the performance on two datasets of λ and η.

TABLE VIII
REPLACING THE FEATURE EXTRACTION NETWORK OF SIAMH ON MARKET-1501 DATASET. THE LENGTH OF HASH CODES IS 128-BITS. "S" DENOTES THE PROPOSED SIAMH

Methods	Market-1501				
	R1 (%)	R5 (%)	R10 (%)	mAP (%)	Q.Time (s)
S&Osnet	88.09	95.12	97.98	76.17	3.3×10^{-2}
S&MGN	82.47	92.30	94.14	69.36	3.3×10^{-2}
S&PCB	86.82	94.11	95.75	75.01	3.3×10^{-2}
S	**90.59**	**96.70**	**98.51**	**78.43**	3.3×10^{-2}

E. Ablation Study

1) Parameter Sensitivity: In this part, we evaluate the influence of λ and η. As shown in Fig.7, the performance of SIAMH is not sensitive to λ. Besides, when $\lambda = 0.1$, SIAMH achieves the best accuracy and Rank-1. As for η, the performance reaches its highest point when η increases to 0.01. Though there exists a slight drop, SIAMH is not sensitive to the value of η overall. Therefore, in all the experiments of SIAMH, we set $\lambda = 1.5$ and $\eta = 0.3$ respectively.

2) Model Flexibility: To verify the model flexibility of SIAMH, we replace the teacher network with existing off-the-shelf real-valued methods, including PCB [1], MGN [8], Osnet [46]. As is revealed in Table VIII, SIAMH can also obtain similar results when combining with these methods, while has no additional time costs. Therefore, compared with existing hashing methods, SIAMH shows great advantages in model flexibility.

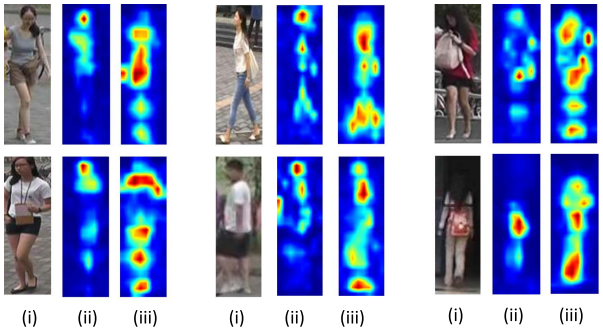


Fig. 8. Activation maps visualization of feature maps. From left to right, (i) Original images, (ii) Activation map of common mutual learning, and (iii) Activation map of IAMT. In the heat map, the response increases from blue to red.

Fig. 9. Retrieval results of SIAMH, CtF, and BoT. The images in the left column are query sets. Blue/red boxes indicate correct/false matches.

3) Qualitative Results: Apart from the accuracy, we visualize the salience and ranked instance retrieval results to provide more convincing insights into SIAMH. In Fig.8, we compare the difference of salience maps from the student model. As the figure shows, the results of common methods are slightly coarse, which are mainly due to two mentioned drawbacks: 1) In the common mutual learning framework, the gap of learning ability between two models does harm to the teacher model, 2) it cannot fully utilize the complementary advantages of two models. To address these issues, IAMT is proposed, in which two model are trained in an asymmetric positive iterative mode. Therefore, compared with the former results, the extracted salience maps of IAMT are more refined, in which the outline of human bodies is more complete. As a result, binary codes are always extracted from the weighted salient regions, thus boosting the performance of SIAMH. Fig.9 shows some instance retrieval results of SIAMH. As it illustrates, SIAMH shows good performance in handling occlusions and distinguishing persons with similar appearance, due to it powerful ability in refining discriminative salience regions.

V. Conclusion

In this paper, we propose a novel deep hashing method for fast person re-identification, called salience-guided iterative asymmetric hashing (SIAMH). Concretely, different from existing works, SIAMH addresses the problem of information redundancy by presenting a new salience-guided self-distillation branch. Moreover, an iterative asymmetric mutual training strategy is developed to alleviate drawbacks of common mutual learning and fully exploit the complementary effect of two models, in which salient regions from the teacher model are continuously refined through iterative training, thus providing a positive circular effect. Extensive experiments on benchmarks show that by explicitly reducing the redundancy of hash codes, SIAMH further closes the gap between real-valued methods and hashing methods in accuracy, whilst significantly reduces the memory cost and response time.

Acknowledgment

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

References

[1] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)* in Lecture Notes in Computer Science, vol. 11220, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Munich, Germany, Sep. 2018, pp. 196–212.

[2] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3356–3365.

[3] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[4] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, vol. 11208, 2018, pp. 501–518.

[5] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3759–3768.

[6] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8295–8302.

[7] X. Zhang *et al.*, "Alignedreid: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: https://arxiv.org/abs/1711.08184

[8] T. Chen *et al.*, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8350–8360.

[9] H. Xiao *et al.*, "Group re-identification: Leveraging and integrating multi-grain information," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 192–200.

[10] J. Chen *et al.*, "Deep local binary coding for person re-identification by delving into the details," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3034–3043.

[11] W. Kang, W. Li, and Z. Zhou, "Column sampling based discrete supervised hashing," in *Proc. 13th AAAI Conf. Artif. Intell.*, D. Schuurmans and M. P. Wellman, Eds. Phoenix, AZ, USA: AAAI Press, 2016, pp. 1230–1236.

[12] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 37–45.

[13] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry (SCG)*, 2004, pp. 253–262.

[14] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1556–1564.

[15] Y. Cao, M. Long, B. Liu, and J. Wang, "Deep Cauchy hashing for Hamming space retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1229–1237.

[16] F. Shen, C. Shen, Q. Shi, A. V. D. Hengel, Z. Tang, and H. T. Shen, "Hashing on nonlinear manifolds," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1839–1851, Jun. 2015.

[17] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2475–2483.

[18] K. Lin, J. Lu, C. Chen, J. Zhou, and M. Sun, "Unsupervised deep learning of compact binary descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1501–1514, Jun. 2019.

[19] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3270–3278.

[20] Y. Zhao, S. Luo, Y. Yang, and M. Song, "DeepSSH: Deep semantic structured hashing for explainable person re-identification," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1653–1657.

[21] V. E. Liong, J. Lu, L.-Y. Duan, and Y.-P. Tan, "Deep variational and structural hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 580–595, Mar. 2020.

[22] Y. Xu, Y. Wang, A. Zhou, W. Lin, and H. Xiong, "Deep neural network compression with single and multiple level quantization," in *Proc. 32nd AAAI Conf. Artif. Intell.*, S. A. McIlraith and K. Q. Weinberger, Eds. Palo Alto, CA, USA: AAAI Press, 2018, pp. 4335–4342.

[23] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao, "Fast person re-identification via cross-camera semantic binary transformation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5330–5339.

[24] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4806–4817, Oct. 2017.

[25] Z. Liu, J. Qin, A. Li, Y. Wang, and L. Van Gool, "Adversarial binary coding for efficient person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 700–705.

[26] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)* in Lecture Notes in Computer Science, vol. 12353, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Glasgow, U.K.: Springer, Aug. 2020, pp. 275–292.

[27] F. Zheng and L. Shao, "Learning cross-view binary identities for fast person re-identification," in *Proc. 25th Int. Joint Conf. Artif. Intell., (IJCAI)*, S. Kambhampati, Ed., 2016, pp. 2399–2406.

[28] D. Chen, J. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proc. 34th AAAI Conf. Artif. Intell., (AAAI)*, 2020, pp. 3430–3437.

[29] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge distillation with adversarial samples supporting decision boundary," in *Proc. 33rd AAAI Conf. Artif. Intell., (AAAI)*, 2019, pp. 3771–3778.

[30] W. Lin *et al.*, "Group re-identification with multi-grained matching and integration," 2019, *arXiv:1905.07108*. [Online]. Available: https://arxiv.org/abs/1905.07108

[31] W. Lin *et al.*, "Learning correspondence structures for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2438–2453, May 2017.

[32] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[33] D. Li, Y. Gong, D. Cheng, W. Shi, X. Tao, and X. Chang, "Consistency-preserving deep hashing for fast person re-identification," *Pattern Recognit.*, vol. 94, pp. 207–217, Oct. 2019.

[34] S. Su, C. Zhang, K. Han, and Y. Tian, "Greedy hash: Towards fast optimization for accurate hash coding in CNN," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst., (NeurIPS)*, 2018, pp. 806–815.

[35] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: https://arxiv.org/abs/1503.02531

[36] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[37] L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng, "Revisit knowledge distillation: A teacher-free framework," 2019, *arXiv:1909.11723*. [Online]. Available: https://arxiv.org/abs/1909.11723

[38] M. Li, Q. Jiang, and W. Li, "Deep multi-index hashing for person re-identification," 2019, *arXiv:1905.10980*. [Online]. Available: https://arxiv.org/abs/1905.10980

[39] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.

[40] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)* in Lecture Notes in Computer Science, vol. 11213, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Munich, Germany: Springer, 2018, pp. 661–678.

[41] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: https://arxiv.org/abs/1610.02984

[42] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[43] Y. Sun *et al.*, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 393–402.

[44] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3183–3192.

[45] X. Chen *et al.*, "Salience-guided cascaded suppression network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3297–3307.

[46] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3701–3711.

[47] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

[48] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[50] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.

[51] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[52] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[53] X. Shu *et al.*, "Large-scale spatio-temporal person re-identification: Algorithm and benchmark," 2021, *arXiv:2105.15076*. [Online]. Available: https://arxiv.org/abs/2105.15076

[54] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "FastReID: A pytorch toolbox for general instance re-identification," 2020, *arXiv:2006.02631*. [Online]. Available: http://arxiv.org/abs/2006.02631

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent., (ICLR)*, Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA, May 2015, pp. 1–15.

[56] X. Wang, H. Zhang, W. Huang, and M. R. Scott, "Cross-batch memory for embedding learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6387–6396.

[57] H. Dong, P. Lu, S. Zhong, C. Liu, Y. Ji, and S. Gong, "Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning," *Neurocomputing*, vol. 307, pp. 25–37, Sep. 2018.

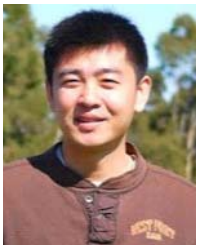
[58] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[59] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.
[60] Z. Cao, M. Long, J. Wang, and P. S. Yu, "HashNet: Deep learning to hash by continuation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5609–5618.

Cairong Zhao received the B.Sc. degree from Jilin University in 2003, the M.Sc. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, and the Ph.D. degree from Nanjing University of Science and Technology in 2011. He is currently a Professor with Tongji University. He is the author of more than 30 scientific papers in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.

Yuanpeng Tu is currently pursuing the master's degree with the College of Electronics and Information Engineering, Tongji University. His research interests include computer vision, deep learning, and person search, in particular, focusing on person re-identification and hashing for visual surveillance.

Zhihui Lai (Member, IEEE) received the B.S. degree in mathematics from South China Normal University in 2002, the M.S. degree from Jinan University in 2007, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology (NUST), China, in 2011. He was a Research Associate, a Postdoctoral Fellow, and a Research Fellow with The Hong Kong Polytechnic University. He has published over 100 scientific articles. His research interests include face recognition, image processing and content-based image retrieval, pattern recognition, compressive sense, and human vision modelization and applications in the fields of intelligent robot research. He is currently an Associate Editor of the *International Journal of Machine Learning and Cybernetics*.



Fumin Shen received the bachelor's degree from Shandong University, Jinan, China, in 2007, and the Ph.D. degree from Nanjing University of Science and Technology, Nanjing, China, in 2014. He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His major research interests include computer vision and machine learning.

Heng Tao Shen (Senior Member, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, in 2000 and 2004, respectively. Then, he joined The University of Queensland and became a Professor in late 2011. He is currently the Dean of the School of Computer Science and Engineering, the Executive Dean of the AI Research Institute, University of Electronic Science and Technology of China (UESTC), and the Dean of Sichuan AI Research Institute (Yibin). His research interests mainly include multimedia search, computer vision, artificial intelligence, and big data management. He is an ACM Fellow and an OSA Fellow.

Duoqian Miao is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University. He serves as the Vice President for the International Rough Set Society (IRSS), an Executive Manager for the Chinese Association for Artificial Intelligence (CAAI), the Chair for the CAAI Granular Computing Knowledge Discovery Technical Committee, a Distinguished Member for Chinese Computer Federation (CCF), and the Vice President for Shanghai Computer Federation and Shanghai Association for Artificial Intelligence. His interests include machine learning, data mining, big data analysis, granular computing, artificial intelligence, and text image processing. He has published more than 200 papers in IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Pattern Recognition*, *Information Sciences*, *Knowledge-Based Systems*, *Chinese Journal of Computers*, *Journal of Software* (in Chinese), *Journal of Computer Research and Development* (in Chinese), *Automatica Sinica* (in Chinese), and *ACTA Electronica Sinica* (in Chinese). He won the Second Prize at Wu Wenjun AI Science and Technology in 2018. He serves as an Associate Editor for the *International Journal of Approximate Reasoning* and an Editor of the *Journal of Computer Research and Development* (in Chinese).